

Chronic Kidney Diseases Detection: A Machine Learning-Based Approach on Bangladeshi Patient Dataset

Tammim Shahara Rafa
Department of EEE
University of Liberal Arts Bangladesh
Dhaka-1207, Bangladesh
tammim.shahara.eee@ulab.edu.bd

Utsab Saha
School of Data and Sciences
BRAC University
Dhaka-1212, Bangladesh
utsab.saha@bracu.ac.bd

Md. Rajib Rayhan
Department of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
rajib2jsr@gmail.com

Sajib Kumar Bosu
Mymensingh Medical College
Mymensingh, Bangladesh
sajibbasu88@gmail.com

Atik Jawad
Department of EEE
University of Liberal Arts Bangladesh
Dhaka-1207, Bangladesh
atik.jawad@ulab.edu.bd

Abstract—Chronic kidney disease (CKD) is a major concern for public health worldwide, impacting around 324 million people globally. It is particularly prevalent in South Asia, with a prevalence rate of 10%. Nevertheless, the deficiencies in healthcare infrastructure in less developed areas hinder the timely detection of early-stage CKD. As the condition worsens, the steady weakening of the immunological and neurological systems requires advanced treatments like dialysis. This study presents an important contribution by introducing a carefully curated dataset of CKD derived from patient reports. By utilizing a proper feature selection technique and SMOTE followed by a range of machine learning methods such as Logistic regression, KNN, SVM, and ensemble method, the proposed predictive model achieves prediction accuracy of 98.03%, with high precision, recall, and F1 score. Significantly, this research is a new initiative in Bangladesh, since it is the first-ever endeavor to generate a dataset for CKD using Bangladeshi patient's data, thereby advancing the field of kidney disease research and diagnostics. Our study presents an important dataset and emphasizes the profound impact of data-driven approaches in healthcare. It provides an extensive resource for advancing research on CKD in Bangladesh and presents an innovative step towards global kidney disease diagnostics.

Index Terms—Chronic Kidney Diseases, Data-driven approach, Machine Learning, Bangladeshi CKD Patient Dataset

disease (CKD) increases the likelihood of developing hypertension, diabetes, and cardiovascular complications, ultimately necessitating dialysis in advanced stages [1], [5]. Numerous factors are taken into account throughout the diagnosis process, such as blood tests and demographics [6]. Major factors that influence chronic kidney disease are depicted in Fig. 1.

For the purpose of CKD prediction, intelligent approaches—in particular, machine learning (ML)—have been investigated in recent time [7]. Dritsas and Trigka showed

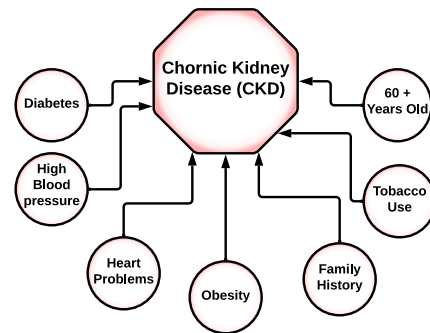


Fig. 1. Factors affecting CKD

I. INTRODUCTION

A global health problem, chronic kidney disease (CKD) affects millions of people each year as a result of genetic predisposition and lifestyle factors [1]. People in less developed areas have obstacles to screening because of financial constraints and limited infrastructure, despite the fact that early identification is essential for decreasing the disease's course [2]. Globally, prevalence rates differ; Sub-Saharan Africa has higher rates, like 13.9%, while West Africa has notable rates like 16% [3], [4]. In addition to urinary issues, chronic kidney

the superior performance of machine learning techniques for chronic kidney disease risk prediction [8]. Arif et. al. introduce an effective machine learning (ML) model for forecasting chronic kidney disease (CKD), which includes several pre-processing procedures, feature selection, a hyperparameter optimization technique, and ML algorithms [9]. A few recent studies also demonstrated that diagnostic accuracy can be improved by feature optimization techniques, deep learning ensembles, and data mining methods [10], [11], [12]. Comparative studies demonstrate the superiority of algorithms; probabilistic neural networks (PNN) are particularly good

at classifying severity stages [13]. Li et. al. developed and validated ML models to predict Mortality in critically ill CKD patients [14]. The XGBoost model is the most successful ML model for assisting clinicians in managing and implementing early therapies in critically ill CKD patients at high risk of death [14]. Rady and Anwar developed efficient data mining techniques using machine learning models like PNN, MLP, SVM, and RBF to accurately identify the severity stage of chronic kidney disease from clinical and laboratory data [15].

After reviewing the existing literature, it is evident that machine learning in CKD prediction is gaining extensive attention. Due to dataset variability—particularly when it comes to CKD identification using blood reports—no one algorithmic approach is sufficient, and further research is required for accuracy. Despite differences in physical characteristics worldwide, Bangladeshi patients are not well represented in current research. A large dataset is necessary for intelligent detection of CKD in Bangladesh, necessitating careful testing of machine learning models in advance of wide implementation. To this end, this paper proposes a new dataset, made up of 101 medical records from Ibn Sina Hospital, for the identification of CKD in Bangladesh. Reports from both affected and unaffected individuals are included, and they are verified by medical professionals. Important features are taken into account, including hemoglobin, white blood cell count, and several blood chemical indicators, etc. The description of each attribute is given in Table I. Following that, a comprehensive analysis of the dataset is performed in this work. Furthermore, a total of three different machine learning algorithms along with their ensemble version are trained on the dataset to analyze the performance of each model in providing accurate CKD prediction results. The main contributions of the paper are-

- This study presents a new dataset for predicting chronic kidney disease (CKD), which is specifically created using data from patients in Bangladesh. This dataset is the first of its kind to be developed in Bangladesh.
- This dataset enables the incorporation of machine-learning techniques to conduct CKD classification.
- The proposed methodology is validated using another benchmark dataset, named CKD Abu Dhabi [16], and obtained superior performance compared to existing state-of-the-art works.

The rest of the paper is organized as follows. Section II describes the data acquisition, data analysis, and prediction of CKD using machine learning models. Section III presents the details of the experimentation and the results obtained. Finally, a brief conclusion of our findings and insights is discussed in section IV.

II. METHODOLOGY

A. Data Acquisition

In this investigative study, we leverage an initial dataset comprising 101 instances, in which 84 patients have CKD and the rest of them are healthy, each delineated by 14 attributes.

Specifically, the quantitative findings of age, white blood cell (WBC), hemoglobin (Hb), poly/neutrophil, mono, lympho, eosino, platelet, blood urea (Urea), creatinine (Creat), sodium (Na+), chloride (Cl-), potassium (K+) and carbon-di-oxide (CO2) are used as the attributes for the CKD dataset. The data is annotated by medical professionals using the standard reference ranges for adult males and females; however, these values may exhibit variation based on factors such as age, gender, and demographic characteristics. The data has been collected from Ibn Sina Diagnostic and Consultation Center, Jessore, Bangladesh [17]. Given the individual demands and stage of CKD of every patient, as well as their medical history and overall health, healthcare providers ought to set more specific goal ranges.

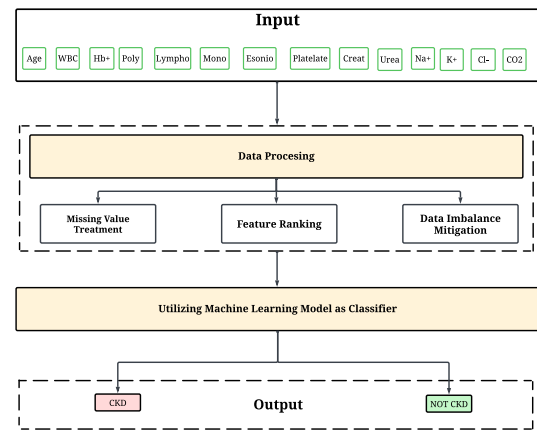


Fig. 2. Flowchart of the proposed algorithm

In Table II, the normal range of each attribute is described. When interpreting these results, careful consideration is given to the specific context of the testing institution. This study involves the collection and analysis of a numerical dataset derived from medical patients from Ibn Sina Diagnostic and Consultation Center, Jessore, Bangladesh. Data preprocessing procedures are employed to address missing values, scale numerical features, and encode categorical variables, ensuring the readiness of the data for subsequent model training and testing. Fig. 2 illustrates the flowchart, detailing the process of dataset generation, data pre-processing, model training, and finally CKD prediction.

B. Data Processing

1) *Missing Value treatment*: The dataset obtained has several missing values for different types of features. Consequently, it is necessary to treat these missing values by replacing them using some relevant values. For this purpose, the random imputation method is utilized to identify missing values, followed by the generation of random values based on the distribution of existing non-missing values. Subsequently, these random values are assigned to the missing entries. Random imputation is known to help prevent bias and ensure the reliability of the data. This method is commonly employed in medical research to ensure that the results are trustworthy.

TABLE I
DATASET DESCRIPTION

Feature	Explanation	Measurement unit	Values
Age	Age of the patient	Integer	[23, 24, ..., 64]
WBC	White Blood Cell the patient's blood	per mm^3	[1160, 12940, ..., 71600]
Hb	Hemoglobin, which is responsible for oxygen transport in the patient's blood	Percentage	[1.2, 8, ..., 12.2]
Poly/Neutrophil	Neutrophil granulocytes of the white cell count in the patient's blood	Percentage	[9, 74, ..., 94]
Mono	Monocytes of the white cell count in blood	Percentage	[0.6, 5.2, ..., 39]
Lympho	Lymphocytes of the white cell count in blood	Percentage	[2, 16, ..., 35]
Eosino	Eosinophil granulocytes of the white cell count in blood	Percentage	[0, 3.9, ..., 19.6]
Platelet	The amount of platelets or thrombocytes in the blood	per mm^3	[14000, 228680, ..., 514000]
Urea	The amount of urea found in the patient's blood	mg/dl	[50.7, 138.72, ..., 427]
Creat	Serum creatinine, Creatinine level in patient muscles	mg/dl	[1.56, 10.40, ..., 34.1]
Na+	Sodium mineral level in the patient's blood	mmol/L	[5.2, 132.57, ..., 146]
K+	Potassium mineral level in the patient's blood	mmol/L	[2.7, 27, ..., 104]
Cl-	Chlorine level in the patient's blood	mmol/L	[7, 98, ..., 121]
CO2	Carbon-Di-oxide level in the patient's blood	mmol/L	[15, 23, ..., 35]
Diagnosis (target)	Diagnosis of chronic kidney diseases (CKD, Not CKD)	Boolean	[0, 1]

TABLE II
NORMAL RANGE OF DIFFERENT FEATURE

Feature	Normal Range
WBC	4.0 to $11.0 \times 10^3/mm^3$
Hb	11.6 and 16.6 gm/dL
Poly/Neutrophil	40% to 75%
Mono	2% to 10%
Lympho	20% to 50%
Eosino	1% to 6%
Platelet	150 to $350 \times 10^3/mm^3$
Urea	15 to 40 mg/dL
Creat	0.57 to 1.26 mg/dL
Na+	135 to 145 mmol/L
K+	3.5 to 5.2 mmol/L
Cl-	96 to 106 mmol/L
CO2	23 to 29 mmol/L

2) *Feature ranking*: In our investigation, the Mann-Whitney U test (MWU) is employed as the primary method for evaluating the significance of continuous features in predicting chronic kidney disease (CKD) events. This decision is driven by the nature of our dataset, which predominantly consisted of continuous variables such as age, laboratory test results, and physiological measurements.

The MWU test, a non-parametric statistical test, is chosen for its several advantages suited to our analytical needs. It is well-suited for assessing differences in the distribution of continuous variables between two groups, making it particularly appropriate for our binary classification task of distinguishing patients with and without CKD events. Additionally, the MWU test does not assume normality in the data distribution, rendering it robust against potential violations of this assumption often encountered in clinical datasets.

Upon applying the MWU test to each continuous feature, relevant statistical metrics, including test statistics, p-values, and effect sizes (Cohen's d), are computed to quantify the strength and significance of associations between features and severe CKD events. These metrics facilitated the identification

of statistically significant predictors while also providing insights into the practical significance of observed differences in feature distributions between patient groups.

3) *Data Imbalance mitigation*: In this section of the analysis, the issue of class imbalance in the dataset is addressed, with a specific focus on mitigating the imbalance between classes in the target variable related to severe chronic kidney disease (CKD) events.

To achieve this, the synthetic minority over-sampling technique (SMOTE), a widely used method for balancing class distributions in imbalanced datasets, is employed. SMOTE works by generating synthetic samples for the minority class, thereby increasing its representation in the dataset and achieving a more balanced distribution between classes. By resampling the data in this manner, it is ensured that the predictive model would be trained on a more representative dataset, thereby reducing the risk of biased predictions towards the majority class. This approach enhances the model's ability to generalize to unseen data and improves its performance in accurately identifying instances of severe CKD events, ultimately contributing to more reliable and actionable insights in clinical decision-making.

C. Classification Using Machine Learning model

In the classification of chronic kidney disease (CKD) and non-CKD instances, three distinct machine learning algorithms, namely K-nearest neighbors (KNN), support vector machine (SVM), and logistic regression, are employed. Leveraging a proposed feature ranking methodology, the most discriminative features are carefully selected. Subsequently, these identified features serve as the basis for distinguishing between CKD and non-CKD cases across the aforementioned classification frameworks. This methodological approach ensures the utilization of the most informative features, thereby enhancing the accuracy and robustness of the classification models in delineating CKD and non-CKD instances.

III. RESULTS AND DISCUSSIONS

The outcomes present the efficacy of different classification techniques on two separate healthcare datasets: one containing patient data from Bangladesh and the other centering on chronic kidney disease (CKD) in the United Arab Emirates. Furthermore, a comparative evaluation is performed with other well-established methodologies utilizing the CKD Abu Dhabi dataset [16].

TABLE III
PERFORMANCE OF PROPOSED METHOD USING BANGLADESHI PATIENT DATASET

Method	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.9411	0.9559	0.9412	0.9443
SVM	0.9607	0.9679	0.9607	0.9626
KNN	0.9215	0.9457	0.9215	0.9267
Ensemble	0.9803	0.9823	0.9803	0.9807

First of all the performance of the proposed method is evaluated using our collected dataset of Bangladeshi patients and results are summarized in Table III. Accuracy, precision, recall, and F-1 Score are used as the performance metrics. From Table III it is quite evident that Logistic Regression and SVM exhibited commendable accuracy rates of 94.11% and 96.07% correspondingly. However, the ensemble method exhibited a more substantial improvement, attaining an impressive accuracy of 98.03%. This finding suggests that ensemble techniques are effective when applied to complex datasets, such as those containing healthcare information, possibly because of their capacity to identify a wide range of patterns and interactions.

TABLE IV
PERFORMANCE CKD ABU DHABI [16] DATASET USING THE PROPOSED METHOD

Method	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.7575	0.7686	0.7575	0.7628
SVM	0.7979	0.8074	0.7979	0.8023
KNN	0.7272	0.7901	0.7272	0.7508
Ensemble	0.8484	0.8269	0.8484	0.8170

For validation of our proposed method, an analysis is conducted on the CKD Abu Dhabi dataset which is shown in Table IV, where logistic regression, SVM, KNN, and the ensemble method are evaluated. Here the ensemble method again demonstrated superior performance compared to all other methods assessed, attaining an accuracy of 84.84%.

Additionally, a comparison of the proposed ensemble method with a few existing works is shown in Table V for the CKD Abu Dhabi dataset. The proposed method demonstrated consistent superiority in terms of accuracy, precision, recall, and F1 scores. This underscores the potential utility of the ensemble method in clinical decision-making and patient care within CKD management protocols, as it demonstrates the

robustness and dependability of the method in identifying CKD cases.

TABLE V
PERFORMANCE COMPARISON WITH OTHER METHOD USING CKD ABU DHABI [16] DATASET

Method	Accuracy	Precision	Recall	F1 score
Random Forest [18]	0.843	-	-	0.55
Neural Network [18]	0.84	-	-	0.35
Proposed Ensemble	0.8484	0.8269	0.8484	0.8170

IV. CONCLUSION

This study underscores the substantial potential of machine learning models for kidney disease detection, as evidenced by the consistently high-performance metrics observed in our proposed dataset. The research not only highlights the promising applications of machine learning in healthcare but also introduces the first-ever chronic kidney disease (CKD) dataset for Bangladeshi patients. The evaluation of this dataset using various machine learning models showcases their efficacy in accurately and efficiently diagnosing kidney diseases, as reflected in the balanced precision and recall values that minimize both false positives and false negatives. While our findings are promising, further validation on diverse and larger datasets is needed to enhance the generalizability of the proposed models. Additionally, future work could explore the integration of additional relevant attributes and consider real-world clinical scenarios to validate the practical utility of the models.

REFERENCES

- [1] N. Bhaskar, M. Suchetha, and N. Y. Philip, "Time series classification-based correlational neural network with bidirectional lstm for automated detection of kidney disease," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4811–4818, 2020.
- [2] J.-C. Lv and L.-X. Zhang, "Prevalence and disease burden of chronic kidney disease," *Renal fibrosis: mechanisms and therapies*, pp. 3–15, 2019.
- [3] M. Davids and M. Chothia, "Chronic kidney disease for the primary care clinician," *South African Family Practice*, vol. 61, no. 5, pp. 19–23, 2019.
- [4] J. W. Stanifer, B. Jing, S. Tolan, N. Helmke, R. Mukerjee, S. Naicker, and U. Patel, "The epidemiology of chronic kidney disease in sub-saharan africa: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 3, pp. e174–e181, 2014.
- [5] S. I. Ali, H. S. M. Bilal, M. Hussain, J. Hussain, F. A. Satti, M. Hussain, G. H. Park, T. Chung, and S. Lee, "Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries," *IEEE Access*, vol. 8, pp. 215 623–215 648, 2020.
- [6] Y. Yang, Z. Tian, M. Song, C. Ma, Z. Ge, and P. Li, "Detecting the critical states of type 2 diabetes mellitus based on degree matrix network entropy by cross-tissue analysis," *Entropy*, vol. 24, no. 9, p. 1249, 2022.
- [7] S. R. Khadka, S. Subedi, B. K. Aidy, L. N. Regmi, A. Parajuli, and M. Bhandari, "Predicting chronic kidney disease using ml algorithms and xai,"
- [8] E. Dritsas and M. Trigka, "Machine learning techniques for chronic kidney disease risk prediction," *Big Data and Cognitive Computing*, vol. 6, no. 3, p. 98, 2022.
- [9] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the early detection of chronic kidney disease: a robust machine learning model," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 144, 2023.

- [10] V. K. Venkatesan, M. T. Ramakrishna, I. Izonin, R. Tkachenko, and M. Havryliuk, "Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease," *Applied Sciences*, vol. 13, no. 5, p. 2885, 2023.
- [11] M. M. Hossain, R. A. Swarna, R. Mostafiz, P. Shaha, L. Y. Pinky, M. M. Rahman, W. Rahman, M. S. Hossain, M. E. Hossain, and M. S. Iqbal, "Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease," *Machine Learning with Applications*, vol. 9, p. 100330, 2022.
- [12] D. M. Alsekait, H. Saleh, L. A. Gabralla, K. Alnowaiser, S. El-Sappagh, R. Sahal, and N. El-Rashidy, "Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models," *Applied Sciences*, vol. 13, no. 6, p. 3937, 2023.
- [13] A. Farjana, F. T. Liza, P. P. Pandit, M. C. Das, M. Hasan, F. Tabassum, and M. H. Hossen, "Predicting chronic kidney disease using machine learning algorithms," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2023, pp. 1267–1271.
- [14] X. Li, Y. Zhu, W. Zhao, R. Shi, Z. Wang, H. Pan, and D. Wang, "Machine learning algorithm to predict the in-hospital mortality in critically ill patients with chronic kidney disease," *Renal Failure*, vol. 45, no. 1, p. 2212790, 2023.
- [15] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019.
- [16] S. Al-Shamsi, D. Regmi, and R. D. Govender, "Chronic kidney disease in patients at high risk of cardiovascular disease in the united arab emirates: A population-based study," *PloS one*, vol. 13, no. 6, p. e0199920, 2018.
- [17] "Ibn sina diagnostic and consultation center," *Jessore, Bangladesh*.
- [18] D. Chicco, C. A. Lovejoy, and L. Oneto, "A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease," *IEEE Access*, vol. 9, pp. 165 132–165 144, 2021.